International Journal of Forecasting 41 (2025) 649–669

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Forecast value added in demand planning☆

Robert Fildes [a,*], Paul Goodwin [b], Shari De Baets [c]

[a] Centre for Marketing Analytics and Forecasting, Lancaster University Management School, United Kingdom
[b] University of Bath, United Kingdom
[c] Open University of the Netherlands, The Netherlands

ARTICLE INFO

Keywords:
Demand planning
Judgmental forecasting
Judgmental adjustment
Sales and operations planning
Forecasting support systems
Efficiency
Bias adjustment

ABSTRACT

Forecast value added (FVA) analysis is commonly used to measure the improved accuracy and bias achieved by judgmentally modifying system forecasts. Assessing the factors that prompt such adjustments, and their effect on forecast performance, is important in demand forecasting and planning. To address these issues, we collected the publicly available data on around 147,000 forecasts from six studies and analysed them using a common framework. Adjustments typically led to improvements in bias and accuracy for only just over half of stock keeping units (SKUs), though there was variation across datasets. Positive adjustments were confirmed as more likely to worsen performance. Negative adjustments typically led to improvements, particularly when they were large. The evidence that forecasters made effective use of relevant information not available to the algorithm was weak. Instead, they appeared to respond to irrelevant cues, or those of less diagnostic value. The key question is how organizations can improve on their current forecasting processes to achieve greater forecast value added. For example, a debiasing procedure applied to adjusted forecasts proved effective at improving forecast performance.

1. Introduction

The judgmental adjustment of forecasts from algorithms and statistical models (hereafter referred to as 'system forecasts', i.e. a forecast made by forecasting software) is ubiquitous, ranging from macroeconomics, weather forecasting, and risk and recommender models to almost all aspects of organizational forecasting. The adjustments are thought to occur primarily to incorporate additional 'soft' qualitative information beyond that incorporated in the system forecast, presuming an improvement in forecast accuracy[1] due to the added value of the new information. The process of measuring the extent to which such judgmental adjustments improve or reduce the accuracy of forecasts, and identifying the circumstances that favour adjustment, is the aim of forecast value added (FVA) analysis (Gilliland, 2008). Studies that have applied FVA have found mixed results, with some indicating forecast improvement under specific circumstances such as promotions (Trapero, Pedregal, Fildes, & Kourentzes, 2013) and high forecaster expertise and low system credibility (Alvarado-Valencia, Barrero, Önkal, & Dennerlein, 2017), and others finding that accuracy is typically reduced (Diamantopoulos & Mathews, 1989; Fildes & Goodwin, 2007; Fildes, Goodwin, Lawrence, &

☆ Representative results based on dataset 1 could be reproduced by CASCaD on the 22nd of November. Other results could be reproduced following a similar process.

* Corresponding author.
E-mail addresses: r.fildes@lancaster.ac.uk (R. Fildes), p.goodwin@bath.ac.uk (P. Goodwin), shari.debaets@ou.nl (S. De Baets).

[1] We normally use the term 'accuracy' to include bias, except where it is important to distinguish between the two.

https://doi.org/10.1016/j.ijforecast.2024.07.006

Nikolopoulos, 2009; Franses & Legerstee, 2009; Mathews & Diamantopoulos, 1986, 1992); see Perera, Hurley, Fahimnia, and Reisi (2019) for a full literature review.

This raises a number of important questions.

1. How likely are adjustments to increase or reduce forecast accuracy?
2. How large are the resulting improvements or reductions in accuracy?
3. What factors are associated with decisions to adjust system forecasts, and how do these factors determine the size of adjustments?
4. In what ways can adjustments damage accuracy and how common are these miscalculations?
5. What are the effects of adjustment direction and size on accuracy?
6. Is it possible to improve accuracy by debiasing judgmental adjustments?

Answers to these questions are important for the design of forecasting support systems and the training of forecasters. They can also have theoretical implications for researchers in the area of human judgment. However, many remain unanswered in the literature, or the findings relating to them have been contradictory. Moreover, research on judgmental adjustment has been based primarily on laboratory experiments (e.g. Carbone, Andersen, Corriveau, & Corson, 1983; De Baets & Harvey, 2020; Goodwin & Fildes, 1999; Sanders, 1992), with relatively few studies being conducted in the field (Perera et al., 2019). These field studies have produced inconsistent results, having adopted different methods and data segmentations.

In this paper we aim to address the above questions by using publicly available data on around 146,000 forecasts from six company-based studies and analysing them using a common framework. We address the contradictory findings by analysing datasets across studies to extract answers independent of the methodological setup and taking into account the variation in data properties. The paper is divided into nine sections. Following a literature review in Section 2, Section 3 discusses the measurement of FVA. Section 4 introduces the available datasets, while Section 5 assesses the effects of judgmental adjustments on forecast accuracy. Section 6 identifies through modelling, factors that are associated with decisions to adjust and the size of adjustments. Section 7 offers explanations as to why some adjustments improve accuracy while others damage it. Section 8 investigates whether debiasing adjustments can improve accuracy. Finally, Section 9 offers a set of conclusions and priorities for further research on this topic that is of both practical and theoretical importance.

2. Literature review

Surveys and in-company studies suggest that the judgmental adjustment of system forecasts is a common practice in companies (e.g. Van den Broeke, De Baets, Vereecke, Baecke, & Vanderheyden, 2019). For example, surveys by Fildes and Goodwin (2007) and Fildes and Petropoulos (2015) found that typically over 37% of system forecasts were adjusted. Field studies give higher values: An in-company study of a pharmaceutical company by Franses and Legerstee (2010) reported an adjustment rate of 90%, while managers in a major international food company studied by Fildes et al. (2009) were adjusting over 90% of their system forecasts. How do these adjustments generally take place? The process of forecast adjustment within demand planning is often organizationally complex, as Fildes and Goodwin (2021) describe in a case study. In essence, a demand planning team designates responsibility for a range of products or SKUs, carrying out a set of tasks (Asimakopoulos, Dix, & Fildes, 2011) to reach a final forecast. In theory, the forecast will be based on information exchanged with others, both within and outside the organization, the features of the forecasting support system (FSS) being used, such as the ease of intervention, and forecasts from a model or algorithm taken from the FSS. The forecaster first decides whether to modify the model-based forecast and, when a modification is judged to be appropriate, a decision is made on the size of the adjustment.

What motivates forecasters to make an adjustment? In a survey by Fildes and Goodwin (2007), managers gave a variety of reasons for making adjustments. Promotional and advertising activity and product price changes were the most common reasons; others included holidays, changes in regulations, insufficient inventories, government policies, competitors' actions, international crises, sporting events, and the weather. In essence, these reasons, whether quantitative or qualitative, consist of information that is perceived to be excluded from the model underlying the system forecast (Lawrence, Goodwin, O'Connor, & Onkal, 2006).

However, decisions to adjust can also reflect a variety of motivational and psychological biases. Company politics and game playing are common motivations for adjustments (Fildes & Hastings, 1994; Galbraith & Merrill, 1996; Mello, 2009; Oliva & Watson, 2009). System forecasts can be adjusted when managers seek to convert them to targets, plans, or decisions. This can be a problem when the difference between these concepts is not acknowledged so that targets, plans, or decisions are treated as valid expectations of future demand (Fahimnia, Arvan, Tan, & Siemsen, 2022; Fildes et al., 2009). In addition, sales forecasts may be adjusted downwards so that managers gain kudos by exceeding their forecasts, or they may be adjusted upwards to make them acceptable to senior managers. Additionally, the predominance of small adjustments in the Fildes et al. (2009) study of four companies suggests that forecasters may be motivated to tinker with forecasts to justify their role.

In addition to motivational biases, there is evidence that psychological biases are widespread in company forecasting (Karelse, 2021). Laboratory studies have revealed a range of cognitive factors that are associated with the propensity to adjust. People tend to see systematic patterns in random movements in time series (Harvey, Ewart, & West, 1997). Because they expect these patterns to continue, they make adjustments to system forecasts that have filtered out the randomness, assuming that the statistical algorithm has failed to register these patterns

(Lawrence et al., 2006). The perceived 'oversight' of the model, and its inevitable errors caused by randomness, can lead to an unmerited distrust in the value of system forecasts and quick abandonment of their use—a phenomenon known as algorithm aversion (Dietvorst, Simmons, & Massey, 2015, 2018). This tendency can be exacerbated because the random movements and associated forecast errors are salient, while the underlying systematic pattern is latent (Kremer, Moritz, & Siemsen, 2011; Massey & Wu, 2005). People also have the ability to develop elaborate explanations (narratives) for random movements, while they are rarely privy to the rationale underlying system forecasts. This can lead to egocentric discounting of the 'advice' embodied in the system forecast (Bonaccio & Dalal, 2006). Taleb (2005) argued that 'statistics are invisible; anecdotes are salient' (p. 112). As a result, unreliable narratives, rumours, and other non-diagnostic information can outweigh the system forecast and lead to its adjustment or replacement (Fildes, Goodwin, & Onkal, 2019).

Effective decisions on whether to adjust a forecast require the identification of the variables (or cues) that are likely to be associated with beneficial and damaging adjustments. Research in the application of fast and frugal heuristics has revealed that cues can be selected because they are salient or available, rather than because they have validity or diagnosticity (Platzer & Bröder, 2012; Platzer, Bröder, & Heck, 2014). These factors, together with others such as limitations of memory, bounded rationality, and unsubstantiated prior beliefs, may also account for the misweighting of valid cues (e.g. Fildes, 1991). To make matters more difficult, experimental evidence has shown that data volatility can disguise diagnostic cues (e.g. Fildes et al., 2019). In studies by Fildes et al. (2019) and Sroginis, Fildes, and Kourentzes (2022), forecasters were easily diverted to use both non-diagnostic soft information and various irrelevant cues in time series patterns. Information like this can be particularly persuasive when it provides support for future outcomes that are desired (such as high sales), while reliable disconfirming evidence may be ignored (Krizan & Windschitl, 2007). This tendency to seek confirming information can result in a bias towards optimism. Additionally, recency bias can cause the most recent error associated with a system forecast to be particularly influential, as forecasters can overreact to it when making a subsequent forecast (Petropoulos, Fildes, & Goodwin, 2016). This over-attention to recent factors may also account for the observation of Van den Broeke et al. (2019), based on forecasters in four companies, that the propensity to adjust and make larger adjustments increased as the forecast lead time became shorter, even though many of the resulting forecasts were made less accurate.

When it comes to the size and direction of adjustments, research has indicated that small adjustments tend to reduce accuracy (Baecke, De Baets, & Vanderheyden, 2017; Fildes et al., 2009) and waste time (Fildes & Goodwin, 2021). Several studies have also found evidence for the damaging effect of positive (or upward) adjustments. In contrast, the less common negative adjustments tend to improve accuracy (Fildes et al., 2009; Syntetos, Nikolopoulos, Boylan, Fildes, & Goodwin, 2009). It is possible that both organizational and psychological factors such as politics and optimism bias encourage positive adjustments, while managers need firm information to be confident in justifying and defending reduced sales forecasts. Additionally, forecasts subject to negative adjustments are bounded at zero, thereby limiting the potential damage to accuracy.

Despite these issues, there is evidence that under the right conditions, judgmental adjustment can improve point forecast accuracy. Such improvements can be expected where humans have important information that is not available to the algorithm underlying the system forecast, such as abnormal conditions (Blattberg & Hoch, 1990) resulting from a forthcoming (Seifert, Siemsen, Hadida, & Eisingerich, 2015) or configural cues (Einhorn, 1974). Researchers in the 'fast and frugal heuristics' field have found that, in some circumstances, humans can outperform sales promotion or price change (Goodwin & Fildes, 1999; Goodwin, Fildes, Lawrence, & Nikolopoulos, 2007; Remus, O'Connor, & Griggs, 1995). Forecasters may also be able to draw on non-linear relationships between cues and outcomes (Seifert et al., 2015) or configural cues (Einhorn, 1974). Researchers in the 'fast and frugal heuristics' field have found that, in some circumstances, humans can outperform statistical models by making simple judgments based on a small number of cues. This is because of the tendency of statistical models to overfit available data, where a small number of cues are sufficient to capture the essence of the signal (Gigerenzer, Todd, & the ABC Research Group, 1999). However, the extent to which judgment can realise its potential and improve accuracy is still an open question, with existing research pointing towards human flaws and inadequate regime change detection (De Baets & Harvey, 2023). In addition, the rise of machine learning algorithms that are designed to take a wider range of factors into account may change the baseline previously occupied by more simple methods.

Although company-based field studies have yielded valuable findings on the role of judgmental adjustment, there is a need for further research. The extent to which the findings of these studies can be generalised is limited by the small number of organisations involved, which have only been observed at a particular point in the evolution of their demand planning systems. For example, the Franses and Legerstee (2009) study involved a single organisation, but with seven distinct business units, while the Fildes et al. study (2009) was based on four companies. Moreover, researchers have recently developed improved methods for measuring accuracy and bias across multiple time series that were unavailable to earlier studies (Davydenko & Fildes, 2013; Davydenko & Goodwin, 2021). Additionally, more insight is needed about the prevalence of causes of reduced accuracy. For example, does this result more often from excessive adjustment or because adjustments are made in the wrong direction? Similarly, to what extent are damaging changes made because forecasters are overreacting to the most recent sales figure or forecast error? Do forecasters tend to give insufficient weight to the system forecast when

combining it with their estimate of the effects of new information? The practical issue in posing these questions is how to improve the systems and processes underlying demand forecasting.

This paper addresses these questions by analysing a large, combined dataset of 147,131 forecasts[2] and actuals obtained from 10 organizations with 22 business units. It extends the work of earlier studies by using the latest accuracy and bias metrics, presenting distributions of FVA, and modelling the factors associated with decisions both to adjust and to determine the size of adjustments. Crucially, it also sets out the benefits from overcoming the use of inappropriate cues in the adjustment decision. As such, it proposes a standardised methodological framework for analysing FVA.

3. Measuring FVA

Measuring forecast accuracy is a controversial subject. It is both arcane and important in that different measures may well produce different results in accuracy comparisons. Despite its deficiencies, the mean absolute percentage error (MAPE), where the error is measured relative to the outcome, remains one of the most used accuracy metrics (Fildes & Goodwin, 2007). However, for the problem we pose, where we wish to compare one method against another, relative error measures are appropriate (e.g. Davydenko & Goodwin, 2021). We use two relative error measures discussed by Davydenko and Fildes (2015). The first is the average relative mean absolute error (AvgRelMAE). This is effective for comparing one method of forecasting, A, here the system forecasts, with the expert adjusted final forecasts, B. Its calculation is as follows: for each SKU, calculate the $MAE_i^A = \frac{1}{n_{Ai}} \sum_{j=1}^{n_{Ai}} Abs\left(error_j^A\right)$ for series i for observation j, using method A over dataset Ai, where n_{Ai} is the number of observations for that series for which there are forecasts from method A. The same calculation is applied for method B. Note that in general, Bi will be the same dataset as A_i. Finally, take the geometric mean of FVA, GMFVA $= \left(\prod_i \frac{MAE_i^B}{MAE_i^A}.\right)^{\frac{1}{n}}$ over the n SKUs. The reasons for using this error measure are three-fold. First, it treats the two alternative methods equally. Second, it is less prone to being affected by outliers compared to a measure based on the arithmetic mean. And finally, its interpretation is straightforward: a ratio of 110% shows that method B (the adjusted forecast) is 10% worse than the statistical benchmark (Davydenko & Fildes, 2013). We subsequently refer to this measure as FVA1.

A second simpler alternative (though with less desirable empirical and statistical characteristics) is for each observation in each SKU, calculate: $r_{AB} = Abs\left(\frac{e_{Bi}}{e_{Ai}}\right)$, where e_{Ai} and e_{Bi} are the errors of methods A and B, respectively, for an individual observation i and then average using the geometric mean (or median).[3] It has the same interpretation. We use this extensively in that it is more flexible when examining small sub-groups of data. Comparing the two measures suggests limited differences. We refer to this measure as FVA2.

Bias, neglected in earlier studies, can be manifested in different forms. Here, we focus on mean bias, which reflects a systematic tendency to forecast too high or low (Davydenko & Goodwin, 2021). Whatever its form, bias and accuracy are not the same things: an adjustment may improve accuracy yet lead to higher bias, and vice versa. We measure relative absolute mean bias for each series (SKU) j as follows:

$$Relative\ Absolute\ Mean\ Error\ (RelAME)_i = \left| \frac{\frac{1}{n_{B_i}} \sum_j^{B_i} e_{BJ}}{\frac{1}{n_{A_i}} \sum_j^{A_i} e_{AJ}} \right|.$$

This is then averaged over SKUs, again using a geometric mean, to give the relative mean bias of method B compared to method A. A value greater than 1 indicates that method B has greater mean bias and vice versa. Additionally, it is of interest to note the percentage of SKUs where either the accuracy and/or bias in the system forecast is improved by judgmental adjustment.

4. Datasets and data selection

Six datasets were used in this study. We consider only one-step-ahead forecasts, although longer horizons are available for some of the datasets (datasets 4, 5, and 6). Each of the datasets includes weekly or monthly statistical systems forecasts; the final forecasts, which may be the same as the system forecast or adjusted; and the corresponding actual outcomes for each SKU. Any observations without such a triple were omitted, as were special circumstances where both the final and the system forecast were both zero. No SKUs were included where there were fewer than six observations, the rationale being that otherwise the demand forecasters would have little experience in deciding their adjustments, if any. Other subsets of the data, for example intermittent demand or new products, may well show different characteristics. The remaining observations make up the cleaned datasets. Note also that the data were trimmed for extreme errors in the final forecast through a 1% trim, again the aim being to exclude the results being dominated by extraordinary circumstances. These constitute the trimmed datasets. Table 1 gives an overview of the datasets and methods used to produce the system forecast by the respective companies.[4]

5. Forecast value added

5.1. How likely are adjustments to increase or reduce forecast accuracy?

Table 2 shows the percentages of SKUs in each dataset where adjustments led to improvements in FVA and/or bias.

[2] This excludes observations where the actual observation and the system and final forecasts are all zero, the cleaned dataset.

[3] Note that in those rare cases where $e_{Ai} = 0$, we replace it by 1.

[4] It should be noted that we did not have any information on individual forecasters. The adjustment behaviour and motivation of the individual forecaster and their consequences are nevertheless an important topic that should be further explored (see Fildes & Goodwin, 2021).

Table 1
Details of the datasets.

Dataset no.	Source	Year of data collection	Company type	Observations	SKUs	System forecast	Further details
1	Fildes et al. (2009)	2003–2005	Three manufacturing companies: 1.1 pharmaceuticals , 1.2 food, and 1.3 household products.	11,296 (10,663 trimmed)	778 (585)	Variants of exponential smoothing.	Monthly
2	Fildes et al. (2009)	2004.02–2005.42	Retail distribution centre for two product groups from different suppliers (2.1 and 2.2).	57,548 (57,293 trimmed)	788 (782)	Variants of exponential smoothing.	Weekly
3	Franses and Legerstee (2009)[a]	Three-year period with start year prior to 2009	Pharmaceutical company.	25,863 (25,863 trimmed)	1101 (1101)	Based on 'techniques such as Box–Jenkins', Holt–Winters' and the like' (p. 37; (Franses & Legerstee, 2009)).	25 Monthly forecasts of sales for 36 countries and seven product groups[b]
4	Van den Broeke et al. (2019)	2014–2015	Subset of the full dataset chosen for compatibility with the other datasets. Three manufacturing companies: 4.1 industrial coating company (with three business units), 4.2 (with two business units), and 4.3 (with three business units) in the B2B and B2C food industry.	52,198 (46,607 trimmed)[c]	2980 (1206)	Moving average for company 4.1, a simple weighted average for company 4.2, and a basic seasonal model for company 4.3.	Company 4.1 made rolling monthly forecasts for every SKU for all relevant regions and different types of customers, leading to multiple forecasts for each SKU. Company 4.2 has a large number of SKUs, specified according to region and production location. Company 4.3 had a limited number of SKUs
5	Kourentzes and Fildes (2023)	2014–2016	Pharmaceutical company.	980 (719 trimmed)	62 (45)	Determined by a forecast selection algorithm, which selected between single exponential smoothing, linear exponential smoothing, and linear trend regression on time.	Monthly forecasts. Longer horizon forecasts are also available
6	Baker (2021)	Start in 2017	Beer manufacturer	6596 (4294 trimmed)	36 (32)	Advanced exponential smoothing-style algorithm.	Weekly beer sales. Includes two sets of adjusted forecasts

[a] Longer lead time forecasts analysed by Franses and Legerstee (2011) are no longer available.
[b] In the individual product group analysis, product group 7 was excluded due to its more limited number of observations.
[c] The discrepancy in dataset size before and after trimming is mostly due to the rule of excluding those cases where both the system and final forecast are equal to 0.

We note that for the majority of company/business units, the overall value-added performance is positive in that forecasts for more than half the SKUs are improved. One company (dataset 6) proved particularly successful at improving its forecasts. However, the retailer (dataset 2) deviates from the overall pattern with only 16.5% of SKUs having an improvement in FVA. Examining business unit-level performance (not shown for brevity), the two units 1.1 and 4.3, which are separate companies with their own processes and product characteristics, proved particularly successful in adding forecast value. However, looking at bias, a different story emerges with dataset 2 (the retailer with the lowest FVA improvement), improving bias for 77% of SKUs. A potential explanation for this discrepancy could lie in a motivation to optimise inventories (see Fildes et al., 2009). Surprisingly, the correlation between the two performance measures was

Table 2
Percentage of SKUs by dataset that show improvements in bias and FVA due to adjustment and the correlation between bias and FVA.

Dataset	No. of SKUs	% SKUs where bias improved	% SKUs where FVA improved	Correlation: Bias improvement and FVA improvement
Set 1	582	40.4	62.4	−0.38
Set 2	759	77.5	16.5	−0.47
Set 3	1100	58.3	49.8	−0.29
Set 4	1188	44.6	55.2	−0.13
Set 5	45	48.9	53.3	−0.33
Set 6	32	37.5	84.4	−0.20
Overall	3697	55.6	51.5	−0.31

N.B. The overall figures are medians across companies/business units.

almost invariably negative, an issue worth exploring in further research, as it is not a mathematical inevitability.

5.2. How large are the improvements or reductions in accuracy and bias?

Table 3 gives a summary of FVA for each of the datasets and the overall FVA (in terms of both accuracy and bias). Error statistics for the system forecast, the final forecast, and a naïve random walk (RW) forecast are also shown. We additionally report the MAPE and MdAPE despite their deficiencies, given their widespread use, and to allow a comparison with other studies. The MAPE and MdMAPE figures are first averaged over each SKU and then the mean or median is taken.

The overall estimates of FVA show that adjustments lead to improvements that are limited (with one noticeable exception). However, as mentioned above, Fildes et al. (2009) found that negative adjustments were more effective than positive ones, with negative adjustments leading to favourable FVA. Van den Broeke et al. (2019) concluded that the results were, however, case-specific. We therefore split the data into positive and negative adjustments as well as giving the overall figures. In calculating the values of FVA, individual SKUs are equally weighted. Because the denominators differ for positive and negative information, the overall figure is not necessarily an average of the two values for the different adjustment directions.

We can derive a number of observations from Table 3 with regard to (1) inter- and intra-company differences, (2) relative bias, and (3) accuracy and error. The table reveals that the datasets differ in their results. The corresponding detailed table in the online supplementary spreadsheet *CoreAnalysis* shows that differences can even be found for different companies or business units within the same dataset. These differences are not surprising, as the datasets cover a myriad of products and differ in the number of SKUs subject to forecasting (e.g. dataset 5 being relatively small). Additionally, the percentage of forecasts that are being adjusted also varies across datasets, ranging from a mere 10.9% in dataset 2, a retailer, to adjustment for nearly *all* forecasts in dataset 3 (97.3%) and dataset 5 (91%), both pharmaceutical companies. Overall, adjustments are made where there is a higher system forecast error, as can be seen by comparing the MAPEs and MdAPEs of the system forecasts for the adjusted and unadjusted forecasts, suggesting that forecasters have some ability to recognise where adjustments are required (Mathews & Diamantopoulos, 1990).

Second, we look at relative bias. Note that this was calculated as the relative absolute mean error per SKU and subsequently averaged over SKUs using a geometric mean. The relative mean bias displayed in Table 3 is a comparison of the system forecast and the adjusted forecast, with values greater than 1 indicating greater mean bias following an adjustment, and vice versa. Overall, negative adjustments were effective at reducing bias, while positive adjustments had little effect. Two extremes are dataset 2, with an exceptionally large mean bias, and dataset 6, with a very low mean bias. Datasets 1 and 4 have lower mean bias through adjustment for both adjustment directions and measured overall. Datasets 3 and 5 are less consistent: both display a larger mean bias for positive adjustments, but a lower one for negative adjustments. Looking at the dataset in its entirety, there is no consistent relationship between the bias and FVA.

Third, we can observe how this translates to the effects of these adjustments on accuracy (FVA1 and FVA2) and error (MAPE and MdAPE). As in previous studies, negative adjustments generally improve FVA, and the benefits can be substantial. In contrast, positive adjustments tend to reduce accuracy. Datasets 1, 4, and 6 show an overall FVA < 1, indicating a beneficial effect of adjustments on accuracy. In dataset 2, there is a deleterious effect of adjustments, with FVA > 1. Appendix A.2 shows the distributions of FVA and associated statistics for four illustrative business units. These can also be used to identify the risks associated with adjustments, in that distributions with long negative FVA tails could prove particularly costly. Although the distributions are all close to log-normal, there are also clear differences between them, particularly in their skew and kurtosis.

6. What factors are associated with decisions to adjust system forecasts?

In this section, we tackle two topics. First, we look at the predictors of when adjustments of system forecasts are made. Then we look at the predictors of the size of adjustments.

6.1. Predictors of when are adjustments made

As we have seen, the literature provides limited evidence as to what prompts demand forecasters to make adjustments. In the current study, no datasets have documented justifications for the adjustments made. However, based on evidence that cues used in decisions are often based on their salience and availability, rather than their diagnosticity (Fildes et al., 2019; Platzer & Bröder, 2012; Platzer et al., 2014), we hypothesise that information displayed in FSSs is likely to prompt adjustments, even when it has little or no predictive value. In addition, forecasters may read too much into recent information such as the latest uplift in demand or the last forecast error when

Table 3
Summary statistics of data, bias. FVA and accuracy for each dataset.

Dataset[a] (total obs.)	Adjustment direction	No. of adj obs.	% (of total obs.)	Relative bias	Relative accuracy: FVA1	Relative accuracy: FVA2	MAPE (MdMAPE) system	MAPE (MdMAPE) RW	MAPE (MdMAPE) final
Set 1 (10,348)	Positive	4117	39.8	0.927	1.019	1.007	33.7 (24.4)	44.4 (32.8)	43.9 (27.5)
	Negative	3229	31.2	0.599	0.805	0.736	89.2 (28.8)	60.6 (31.3)	34.4 (21.5)
	Overall adjusted	7346	71.0	0.742	0.907	0.891	60.8 (26.9)	49.4 (32.0)	41.1 (41.1)
	Unadjusted	3002	29.0	n.r	n.r.	n.r.	30.0 (20.5)	45.5 (31.2)	n.r.
Note	Two of the three companies show no value added for positive adjustments and with limited improvement in bias.								
Set 2 (55,558)	Positive	2944	5.3	2.106	1.664	1.961	32.6 (27.3)	28.4 (25.0)	57.1 (48.4)
	Negative	3091	5.6	1.753	1.537	1.485	42.9 (26.3)	29.2 (21.2)	44.9 (38.0)
	Overall	6035	10.9	2.230	1.614	1.678	29.1 (51.4)	38.0 (29.1)	51.4 38.0
	Unadjusted	49 523	89.1	n.r.	n.r.	n.r.	25.2 (19.8)	20.7 (17.6)	n.r.
Note	Both suppliers to the retailer fail to add value or lower bias. Fildes et al. (2009) speculated that there was motivational confusion between the forecast and the inventory/service-level decision. The system forecast was worse than the random walk!								
Set 3 (25,345)	Positive	14 413	56.9	1.366	1.133	1.235	32.3 (25.7)	42.7 (29.6)	47.3 (29.2)
	Negative	10 247	40.4	0.675	0.823	0.837	65.6 (31.4)	59.6 (23.1)	35.3 (24.4)
	Overall	24 660	97.3	1.294	1.006	1.043	49.7 (29.3)	52.1 (30.2)	41.4 (26.8)
	Unadjusted	685	2.7	n.r.	n.r.	n.r.	39.9 (31.9)	48.9 (31.0)	n.r.
Note	All the business units show the same pattern.								
Set 4 (45,596)	Positive	22 080	48.4	0.691	1.048	1.074	178.0 (66.6)	no lagged data available	355.5 (89.1)
	Negative	20 193	44.3	0.587	0.698	0.784	384.9 (93.7)		218.6 (68.9)
	Overall	42 273	92.7	0.616	0.613	0.965	339.1 (75.9)		293.9 (77.6)
	Unadjusted	3323	7.3	n.r.	n.r.	n.r.	211.5 (54.1)		n.r.
Note	Companies 1 and 2 both show added value for negative information but not positive. The bias result is inconsistent across companies. Company 3, with only 572 observations, all of which were adjusted, shows great improvement in both bias and FVA, whatever the direction of adjustment.								
Set 5 (654)	Positive	404	61.8	1.077	1.139	1.081	52.7 (50.0)	209.8 (57.6)	152.2 (66.7)
	Negative	191	29.2	0.726	0.922	0.873	204.4 (37.4)	139.0 (48.4)	41.1 (38.7)
	Overall	595	91.0	0.922	1.085	1.009	116.2 (47.3)	180.1 (50.5)	105.6 (50.9)
	Unadjusted	59	9.0	.n.r	.n.r	.n.r	51.2 (38.0)	90.8 (51.7)	.n.r
Set 6 (3911)	Positive	1807	46.2	0.213	0.466	0.502	63.3 (53.2)	35.9 (30.6)	45.5 (40.3)
	Negative	1953	49.9	0.490	0.814	0.745	81.8 (47.5)	41.6 (24.4)	38.0 (29.8)
	Overall	3760	96.1	0.287	0.511	0.616	72.2 (52.0)	38.6 (27.4)	41.9 (33.1)
	Unadjusted	151	3.9	n.r.	n.r.	n.r.	41.8 (20.3)	35.4 (18.4)	n.r
Overall (141,412)	Positive	46 151	32.3	1.002	1.088	1.118	43.2 (38.6)	42.7 (30.6)	51.3 (44.3)
	Negative	38 904	30.2	0.637	0.819	0.811	85.5 (34.4)	59.6 (24.4)	39.5 (33.9)
	Overall	84 669	59.9	0.832	0.957	0.987	66.5 (49.3)	49.4 (30.2)	46.7 (39.5)

(*continued on next page*)

Table 3 (*continued*).

Dataset[a] (total obs.)	Adjustment direction	No. of adj obs.	% (of total obs.)	Relative bias	Relative accuracy: FVA1	Relative accuracy: FVA2	MAPE (MdMAPE) system	MAPE (MdMAPE) RW	MAPE (MdMAPE) final
	Unadjusted	56 743	40.1	n.r.	n.r.	n.r.	40.9 (26.2)	45.5 (31.0)	n.r.

N.B. The overall figures for accuracy include just the adjusted observations. The values on FVA and accuracy are geometric means[b] across companies/business units. The MAPE is measured by the mean of the MAPEs calculated for each business unit/company, while the median values are medians of the corresponding MAPEs. 'n.r.' denotes cells that are necessarily empty.

[a] The number of observations in the cleaned and trimmed data. The number of observations is larger than in Table A.1 where observations are lost due to lagged transformations.

[b] In general, we use medians across companies/business units rather than an overall calculation in order not to give undue weight to an individual company or unit, as is the case in dataset 4, company 4.2, for example.

Table 4
Variables considered in the model and the rationale for their inclusion.

Variable	Rationale	Measure
Previous direction of adjustment	Behavioural inertia; knowledge of the particular SKU's characteristics.	$sign(FFC_{t-1}\text{-}SFC_{t-1})$
Change	Visual stimulus in graphs. Larger changes may lead to a higher probability of adjustment	$SFC_t\text{-}A_{t-1}$
Current system forecast	Visual stimulus.	SFC_t
Previous system forecast error	May be alerted in software, visual stimulus.	$A_{t-1}\text{-}SFC_{t-1}$
Previous final forecast error	May be alerted in software, visual stimulus.	$A_{t-1}\text{-}FFC_{t-1}$
Unobserved information	Demand forecasters typically gather information in the planning process that is not part of the statistical model but is seen as diagnostic.	$A_t\text{-}SFC_t$: positive (*DIPos*) and negative information (*DINeg*) may be responded to differentially.
Extreme last error (final and system)	Large errors are salient.	The size of the last error based on: $Abs(A_{t-1}\text{-}SFC_{t-1})/StdSFC$ or $Abs(A_{t-1}\text{-}FFC_{t-1})/StdSFC$.
Extreme change	Only large changes may be salient and make adjustment more likely.	Size of change: $Abs(A_{t-1}\text{-}SFC_t)/StdSFC>2$.
Extreme (unobserved) information)	Adjustments are primarily made when there is strong new information. Adjustments are more likely in the direction of the extreme information; less likely in the opposite direction.	Size of $abs(A_t\text{-}SFC_t)/StdSFC$ split into 3 classes , less than 1 stdSFC, between 1 and 2 and > 2 *StdSFC*.
SKU-specific features	Demand for some SKUs is poorly characterised by the statistical model compared to others.	Dummy variable for each SKU.

Notation: A_t = actual at time t; SFC_t = system (or statistical) forecast for period t, FFC_t = final (adjusted) forecast at t; *StdSFC* = standard deviation of system forecasts. *DIPos* > 0 when *Actual* > *SFC*, and otherwise 0, while *DINeg* < 0 when *Actual* < *SFC*, and otherwise 0.

these merely reflect noise (e.g. Goodwin & Fildes, 1999; Kremer et al., 2011; Lawrence et al., 2006; Massey & Wu, 2005; Petropoulos et al., 2016).

We therefore define six variables that *may* be of importance for adjustment behaviour (see Table 4 for an overview): (1) previous direction of adjustment, (2) change between the latest forecast and the latest observation, (3) latest system forecast, (4) last error in the system forecast, (5) last error in the final forecast, and finally, (6) unobserved information about forthcoming events (we refer here to information excluded from the computer model that accounts for the system forecast error; some elements of this may be observed by the forecaster and motivate them to making an adjustment). In addition, we consider where the decision to adjust is based on non-linear extreme cases: extreme last error (of the system and final forecast), extreme change, and extreme unobserved information. Other variables might have proved relevant, and have been left for subsequent analyses. These include the absolute error and the previous seasonal adjustment, though the software is unlikely to have made them particularly salient, and it is unlikely that they would have influenced the substantive results we present.

In the discussion below, SFC_t is the statistical (or system) forecast for period t, A_t is the actual outcome for t, and FFC_t is the final forecast for t.

First, it is likely that the direction of the previous adjustment $sign(FFC_{t-1}-SFC_{t-1})$ plays a role in predicting the subsequent adjustment. There can be a variety of reasons for this: behavioural inertia (Gal, 2006), anticipating an event which is now overdue, knowledge of the particular SKU's characteristics (Lim & O'Connor, 1996), persistent optimism (Fildes et al., 2009), or organizational influences (Webby & O'Connor, 1996). Second, change ($SFC_t\text{-}A_{t-1}$), where a positive value indicates that the most recent system forecast is above the previous actual, might be taken to mean that there is an upward change in the level of the series (and vice versa)—or alternatively, that the system forecasts are out of line with the actuals (Andreassen & Kraus, 1990). Third, the latest system forecast itself (SFC_t) may prove to be a visual stimulus for adjustment (Jarvenpaa, 1990), though there is evidence that when special events are forthcoming it may be ignored and replaced by a purely judgmental forecast (Goodwin & Fildes, 1999). Fourth, the error of the last system forecast ($A_{t-1}-SFC_{t-1}$) was found to have an effect by Mathews

and Diamantopoulos (1990) in their early research. They found that the products selected for adjustment typically had larger system forecast errors ($A_t - SFC_t$). This last observed error may be flagged by the software and provide a visual stimulus for change. Fifth, the error in the previous final forecast ($A_{t-1} - FFC_{t-1}$) may also influence the forecaster's propensity to make an adjustment, as it provides outcome feedback on the effectiveness of their most recent intervention. Sixth, adjustments may be made due to unobserved information—that is, information available to the forecaster but not included in calculating the system forecast—that may at least in part account for the system forecast error. Recall that possession of this information is the primary justification for adjusting system forecasts. We use the system forecast error, $A_t - SFC_t$, as a proxy for this information, though it may partly reflect random variation. It is a direct measure of the potential information beyond that already included in the system forecast. Positive information (suggesting an upwards adjustment to a forecast was necessary), such as a forthcoming sales promotion, may well be interpreted differently than negative information (such as a forthcoming increase in tax on a product), so the proxy variable, $A_t - SFC_t$, needs to be split between the positive and negative. Rational expectations arguments, where the demand planner is presumed to have knowledge of any forthcoming events apart from random noise, suggest that the forecaster's final forecast will coincide with the noise-free (i.e. the expected) outcome.

In addition to the six variables described above, extreme cases such as an extreme last error in the system or final forecast, an extreme change, and extreme unobserved information may elicit adjustment due to their high saliency (Andreassen & Kraus, 1990). The extremity of the previous final forecast error was an important determinant of the next adjustment in a study by Petropoulos et al. (2016), the source of dataset 3 in this paper.

Additional variables that are undoubtedly relevant for any particular SKU and have been considered in earlier research include the type of product, such as whether it is perishable (Khosrowabadi, Hoberg, & Imdahl, 2022); the value of the product, using for example, an ABC-XYZ classification; and the individual characteristics of the forecaster, such as their level of experience in the task and traits such as openness to experience, extraversion, cognitive reflection, and external locus of control (Eroglu & Croxton, 2010; Eroglu & Sanders, 2021; Moritz, Siemsen, & Kremer, 2014). The datasets do not in general include these additional variables, and in our modelling we therefore use an SKU-specific categorical variable to capture these effects.

The 'core' variables described above are displayed in Table 4 together with a number of additional potential stimuli that are considered for the reasons indicated. We note where there is evidence of likely impact.

The multinomial logit model used to estimate the effects of the variables shown in Table 4 on the probability of making an adjustment is shown below:

Logit (adjustment) = f (Previous direction of adjustment, Change, Current system forecast, Previous system forecast error, Previous final forecast error, Unobserved diagnostic info positive, Unobserved diagnostic info negative);

where adjustment takes on the value $+1$ for a positive adjustment, -1 for negative, and 0 for no adjustment. The continuous variables are all normalised by the system forecast standard deviation. There are two reasons for this: the first is statistical, in that it removes heterogeneity in the (regression) model; and second, the stimuli are all likely to be interpreted relative to the core variation in the data (rather than the variation, as it is affected by outliers caused by special events).

The approach we adopted is to present the results for the individual companies (or in the case of dataset 3, business units) using a basic model for all datasets, including SKU categorical variables, whether or not they are significant overall. Note that we do not explicitly include the variables measuring extremes. In running the analyses both with and without extreme values, we found limited differences in the general results, either within datasets, where classification accuracy is hardly improved, or across datasets. Nevertheless, we comment on any effects from the inclusion of extremes in Appendix A.1. Where there is more than one 'company' in the dataset, medians across the units in each dataset were used in the table to calculate the summary statistics measuring impact. An observation is classified into one of the three classes (adjust up, no adjustment, adjust down) depending on which has the highest estimated probability. The classification results shown are for the percentage of successful predictions of whether an adjustment took place and its direction. These results are compared with the estimates from our cross-validation exercise, a 10-fold replication of estimators derived from using approximately 75% of observations as a training dataset and 25% as test data. We also examined the effects of including extreme measures on the classification accuracy.

Table 5 summarises the results of the logistic regression in order to identify common features with the full business unit/company details shown in the online material. The impact values greater than 1 indicate a positive association between the variable and the propensity to adjust, and values smaller than 1 indicate a damping effect. Dataset 4 is excluded because it lacks lagged variables.

The impact values were calculated as follows. First, the median coefficient was obtained for the business units or companies within a dataset. Then the median across the datasets was calculated (i.e. the median of the medians). The reason for this choice of summary statistic is to ensure that no single business unit or dataset is given too much weight in the summary statistics. The table shows the 20–80 percentile range for the coefficients as estimated for each dataset, excluding the two extreme estimates.

An impact is described as consistent across datasets if all estimates are greater (less) than or equal to one. With the definitions of the different variables, particularly the 'change' and the 'system forecast' variables, it might be expected that the estimates would suffer from multicollinearity, although many of the datasets are large. With logit models and many dummy SKU variables, no

Table 5
Variables used in the model and their estimated impact on the probability of making an adjustment: Five trimmed datasets.

Variable	Measure	Direction of adjustment		Impact: Median [20% & 80% percentiles]	Comment
Previous direction of adjustment	FFC_{t-1}-SFC_{t-1}	Previous up	Up	2.27 [1.36, 5.28]	Consistent upward effect: greater upward effect if previous adjustment was also up rather than down, similarly with downward adjustments for 4 out of 5 datasets.
			Down	1.52 [0.59, 2.44]	
		Previous down	Up	2.70 [1.34, 4.71]	Consistent downward effect: greater downward effect than when previous adjustment was up, i.e. 4.87>1.52.
			Down	4.87 [2.75, 9.35]	
Change	SFC_t-$A_{t-1:}$		Up	0.93 [0.80,1.03]	Inconsistent.
			Down	1.58 [1.09, 2.64]	Consistent for 4/5 sets.
System forecast	SFC_t		Up	0.74 [0.72, 0.83]	Consistently < 1.
			Down	1.13 [1.05,1.30]	Consistent for 4/5.
Last error: system	A_{t-1}-SFC_{t-1}		Up	0.97 [0.92, 2.24]	Inconsistent.
			Down	1.17 [1.16,1.77]	Consistently >1 for 4/5. Counterintuitively, a positive last error increases the probability of a downward adjustment.
Last error: final	A_{t-1}-FFC_{t-1}:		Up	0.81 [0.35, 0.85]	Consistently <1.
			Down	1.09 [0.06, 1.21]	Consistent apart from dataset 5.
Unobserved information	A_t-SFC_t if > 0: positive (DIPos)		Up	1.22 [1.17, 1.43]	Consistently >1 for 4/5.
			Down	0.90 [0.89, 0.99]	Consistently <1 for 4/5.
	A_t-SFC_t if < 0: negative (DINeg)		Up	0.90 [0.70, 1.29]	Inconsistent.
			Down	0.64 [0.50, 0.81]	Consistently <1, 4/5, i.e. large negative information makes downward adjustment more likely.
Classification accuracy		Whole sample :		80.3%	Test data.
			Up	82.3%	Up 82.1%
			Down	71.4%	Down 66.8%

Notation: A_t = actual at time t; SFC_t = system (or statistical) forecast at t, FFC_t = final (adjusted) forecast at t; *StdSFC* = standard deviation of system forecasts. *DIPos* > 0 when the *Actual* > *SFC*, and otherwise 0, while *DINeg* < 0 when *Actual* < *SFC*, and otherwise 0.

simple statistics are available. Instead, we examined the robustness of the coefficient estimates again using the 10-fold replication employed previously. This judgment of consistency is based on the simulated estimates scored for consistency in their signs (i.e. <1 or >1 when exponentiated), as detailed in the online spreadsheet *Adjust*

(in worksheet *Summary*). Overall, the signs of the simulated coefficients are highly consistent within datasets, with the exception of the signs for 'change' and 'system forecast' for the two smaller datasets. Across datasets, as Table 5 shows, there is again a high level of agreement as to the direction of impact of the different variables—though not all, e.g. the last system forecast error. It is doubtful whether consistency across datasets is to be expected given that the results describe the process, data, and statistical methods from different companies: it is noteworthy when it is observed.

A number of conclusions can be drawn from Table 5 for the six investigated variables, and these are not affected by the limited multicollinearity observed. First, a previous adjustment makes a subsequent adjustment more likely. While its direction is the most important influence, the very fact an adjustment was made itself proved significant in increasing the likelihood of a subsequent adjustment. Second, looking at 'change', the expected effect is present, but depends on the direction of adjustment: a larger positive change (i.e. a larger positive discrepancy between the system forecast and the last actual) makes a subsequent positive adjustment less likely, while a downward adjustment becomes more likely. Third, the system forecast in itself has a consistent effect, such that a higher system forecast makes an upward adjustment less likely and a downward one more likely. Looking at the errors, a system forecast error is not consistent across datasets. The previous final forecast error has an effect that again differs according to direction of the adjustments: a large previous final forecast error leads to a larger likelihood of downward adjustments, but a lesser likelihood of upward adjustments, counter-intuitively.

Finally, the unobserved information has a relatively minor impact on the decision and direction to adjust, despite its theoretical importance, in that all the weight should be placed on the positive information (*DIPos*) for a positive adjustment and vice versa. From a theoretical perspective, forecasters should be adjusting only relating to the diagnostic information, and no weight should be assigned to the other cue variables. The high impact of the non-diagnostic cue variables on the forecaster's judgment suggests that the FSS and the associated processes, complex though they seem to be (see e.g. Fildes & Goodwin, 2021), are not particularly effective at influencing the decision to adjust, based on valid cues only. The analysis suggests that current FSS design leads to a focus on visual artefacts and irrelevant figures.

While the estimates from the individual datasets are broadly consistent, within the datasets we see differences. For example, dataset 5 (pharmaceuticals) shows different variables being more impactful. For some variables, i.e. 'change', 'system forecast', and 'final forecast error', the median estimates across datasets are tightly estimated (as measured by the 20–80 percentile range). The unobserved diagnostic information is such that where this information is positive (*DIPos*), an upward adjustment, as expected, is consistently more likely, though the impact is limited (+22%). For negative information (*DINeg*), again a downward adjustment is more likely (55%). Both effects are tightly estimated. As support for the validity of these models, the classification results appear robust. Perhaps surprisingly, this provides evidence that negative information is more salient (1.55) than positive information (1.22) in affecting the adjustment decision. As the replications show, overall, the key conclusions are robust (details are found in the online spreadsheet *Adjust*).

In conclusion, we see that the basic model is successful overall at classifying the adjustment decisions (apart from those cases where there are few observations in a particular class, such as the beer manufacturer in dataset 6 where almost all observations are adjusted). Critically, consistent effects were identified, namely those of the previous direction of adjustment, change, and the latest error in the final forecast.

6.2. Predictors of the size of adjustments

Qualitatively, the overall model we examine based on the cues we identified in Table 4 is as follows:

Adjust model: $(Adjustment)_t = f(Actual_{t-1}$; *Previous adjustment*; $System\_Forecast_{t-1}$; *Change between current forecast and* A_{t-1}; $Previous\ forecast\ error_{t-1}$; *Unobserved information*).

The earlier analysis showed the potential importance of discriminating between positive and negative adjustments. This suggests a model in which the parameters depend on the direction of adjustment. It is equally likely that the parameters depend on the company or business unit, capturing the organisational aspects of the forecasting process. We comment on any observed difference whilst presenting an overall explanation of the adjustment process. Company/business unit dummies are included to partially capture these differences. Too few observations are available to warrant the inclusion of potential individual SKU effects.

The modelling procedure within each dataset is as follows.

1. Model individual cleaned trimmed datasets by adjustment direction with class variables for companies/
business units.
2. The specification of the model is designed to be robust across datasets.[9] Various specifications were examined with the dependent variable, log(absolute (normalised_adjustment)), being chosen for its better distributional properties (the normalization uses the standard deviation of the system forecast). All other variables were also normalised using the standard deviation of the system forecast, as justified in Section 6.1, and included linearly.
3. Identify influential observations and outliers—removing these from the datasets being analysed.
4. Model overall and by company for the modified dataset by adjustment direction, including SKU effects.

[9] This necessarily implies that the chosen model is not optimal. No major discrepancies from the individual best model were found apart from those noted.

Table 6
Median regression coefficients in a model of log abs(normalised_adjustment): All data.

Data set		Mean: nadj	Previous Adjustment	System forecast	Change	Previous final forecast error	Positive info DIPos $_t$	Negative info DINeg $_t$	% variation	RSq Adj
(no obs. used)		(Median)	Adj $_t$	SFC_t	(SFC_t- A_{t-1})	A_{t-1}-FFC_{t-1}	A_t-SFC_t>0	A_t-SFC_t<0		
Set 1	Up	1.17	0.018	−0.029	−0.049	−0.031	**0.171**	0.045	79.4	51.6%
3186		(0.71)	0.022	−0.018	−0.027	0.003	0.170	0.052		
Set 2		1.00	0.008	−0.029	**−0.778**	**−0.686**	0.096	−0.095	6.7	67.8%
2550		(0.69)	−0.019	−0.034	−0.792	−0.694	0.085	−0.091		
Set 3		1.16	−0.046	**−0.119**	**−0.351**	**−0.332**	0.082	−0.013	4.0	48.8%
12575		(0.75)	−0.046	−0.107	−0.375	−0.355	0.084	−0.014		
Set 5		1.07	0.067	0.019	**−0.230**	*−0.180*	0.020	0.022	1.7	37.5%
285		(0.84)	0.156	0.093	−0.053	−0.014	0.008	0.041		
Set 6		1.16	**0.695**	**−0.231**	0.031	*0.152*	**0.409**	**−0.503**	4.4	74.0%
1640		(0.60)	0.687	−0.197	0.094	0.162	0.392	−0.301		
Set 1	Down	−0.72	−0.124	**0.237**	0.021	−0.083	0.082	**−0.367**	27.5	56.7%
2474		(−0.51)	−0.135	0.223	0.055	−0.041	0.008	−0.366		
Set 2		−1.57	−0.144	**0.494**	**0.236**	0.302	0.162	**0.049**	6.4	21.6%
2698		(−0.83)	−0.137	0.414	0.306	0.364	0.112	0.009		
Set 3		−0.85	0.025	**0.272**	**0.234**	**0.220**	0.046	**−0.122**	4.0	48.7%
8550		(−0.61)	0.035	0.259	0.268	0.260	0.029	−0.121		
Set 5		−0.54	0.074	−0.080	*0.198*	0.143	0.036	**−0.330**	75.2	52.0%
92		(−0.48)	0.105	−0.087	0.220	0.105	0.013	−0.320		
Set 6		−0.46	**−0.751**	0.028	**0.335**	**0.382**	0.155	**−0.420**	9.1	66.2%
1717		(−0.30)	−0.632	0.020	0.370	0.365	0.119	−0.394		
Median	Up		0.018	−0.029	−0.230	−0.180	0.096	−0.013	4.354	51.6%
				Consistent	Consistent	Consistent	Consistent			
Overdata sets	Down		−0.140	0.216	0.234	0.220	0.046	−0.330	9.053	52.0%
			Consistent	Consistent	Consistent	Consistent	Consistent			

N.B. The number of observations shown are those available for use. The actual numbers used in the individual regressions are shown in the online spreadsheet *Regmodels*. The last four rows show medians over data sets and are noted as consistent if at least four of the estimated signs are the same.

5. Identify major differences by company/business unit.
6. Carry out a 10-fold replication on a split of 75% training, 25% testing to check the robustness of the parameter estimates.

The aim of this model is to examine the weightings given to the system forecast and the obviously salient cues. These are (1) the change between the last observed actual A_{t-1} and the latest system forecast for period t, (2) SFC_t, the system forecast on its own, and (3) the previous final forecast error. This last variable is moot, in that in many if not all FSSs, it is not brought to the forecaster's attention except in the situation where the error is large and outside any control limits. However, it should be noted that Petropoulos et al. (2016) showed this to be relevant in the case of large errors, raising the question of whether the effect is non-linear. This is explored below.

The variables included are therefore effectively SFC_t, A_{t-1}, FFC_{t-1}, and SFC_{t-1}, as well as unobserved and extreme unobserved diagnostic information. Categorical effects for the business unit/company were also included but not reported here, as the quantitative differences between them are not within the scope of this analysis. There is potential collinearity between variables, in particular for 'change' and 'previous forecast error'. This was evaluated through the detailed simulation results given in the online output *Regmodels,* as well as in a supplementary regression with no shortage of observations (but without SKU effects). Table 6 shows the resulting models for each of the datasets except dataset 4, which was again excluded because of the lack of available information on lagged variables.

The percentage of variation explained by the unobserved information measured by the additional variation explained/total variation excluding SKUs (SS type I) is also shown. This measures the relative importance of the unobserved information compared with the cue variables. Formal significance tests were carried out (shown in the online supplement). Parameter estimates are shown in bold where a p-value of less than 0.01 holds for over half the companies/business units: an estimated effect of 0.1 on the percentage normalised adjustment of $(e^{0.1}-1)\times 100 \approx 10\%$). Variables that do not meet this criterion are shown in the regular font. The medians of the replications are shown in light grey.

Table 6 shows that all the major variables have an impact, though the specifics depend on the individual datasets. Inevitably, there is variation across datasets, but there are also key consistencies. In the models of the absolute size of the adjustment, the key findings are listed below.

- Over all the datasets modelled, the unobserved diagnostic information has a relatively weak impact, particularly for datasets 2, 3, and 6. Broadly, 'negative information' has the larger impact on the size of the absolute adjustment; the more negative the information the larger the absolute adjustment. Note that a negative coefficient multiplied by negative

information indicates a positive association in the size of the absolute adjustment.
- The latest system forecast also has a broadly consistent effect (with the exception of dataset 5). A larger system forecast delivers a larger adjustment though this is moderated by change, ($SFC_t - A_{t-1}$), which consistently shows that a 'change', where the system forecast deviates positively from the last actual, will dampen an upward adjustment and increase a downward adjustment. A change where the system forecast deviates negatively from the last actual will amplify an upwards adjustment and decrease a downwards adjustment. In both cases, this suggests a tendency to adjust the system forecast to a more 'moderate' level.
- The change variable also proves consistent: when large increases in demand were predicted, forecasters' upward adjustments tended to be smaller, while downwards adjustments were larger. Larger predicted decreases in demand were associated with larger upward adjustments and smaller downward ones.
- The responses to the previous forecast error are broadly consistent,
- There are inevitable differences between companies/business units within datasets: here we have emphasised consistency.

7. In what ways can adjustments damage accuracy and how common are these miscalculations?

A judgmental adjustment to a forecast can improve accuracy and hence lead to a beneficial FVA (i.e. an FVA < 1) if it meets two conditions: it must be in the right direction, and it must not be excessive. We define an excessive adjustment as an adjustment in the right direction where the absolute error of the final forecast exceeds twice the absolute error of the system forecast. For example, a final forecast of 141 would be excessive if the system forecast was 100 and the actual demand was 120. In addition, we define an optimistic adjustment as one where the final forecast exceeds the actual demand.

7.1. Effects of excessive and wrong-direction adjustments

The decision tree in Fig. 1 shows where a decision to adjust upwards or downwards will be beneficial (i.e. FVA < 1) or damaging (i.e. FVA > 1). Also shown are the median percentage of adjustments across all 16 business units that belonged to each decision or outcome, and the median percentage of adjustments for each pathway, together with the resulting FVA for accuracy and bias.

It can be seen that 64% (100% – 22.5%–13.5%) of forecasts were in the right direction, and 56.1% of the overall adjustments were upwards. The majority of these were optimistic (67.3%). Four pathways through the tree result in beneficial adjustments (shown in bold lines), and the FVA figures for accuracy (0.414 to 0.247) show that these typically reduced errors by around 60% to 75%. Non-excessive downward adjustments in the right direction were most effective. The four pathways that result in damaging adjustments reveal that excessive adjustments, although rarer than those in the wrong direction, were most damaging. They typically increased errors by around 3.5 times, while wrong-direction adjustments increased errors by 2 to 2.5 times. Excessively optimistic positive adjustments that were in the right direction proved to be particularly damaging. The failure to achieve positive FVA on average is because the excessive and wrong-direction adjustments increased errors by 174%, while beneficial adjustments—many of them modest—only improved errors on average by 61%.

The consistency across business units is also important. Table 7 shows, alongside the medians,the interquartile ranges across the units for the percentage of adjustments on each pathway in the tree and the FVA.The small IQRs suggest a high degree of consistency across the businesses.

7.2. What are the effects of adjustment size on accuracy?

Fildes et al. (2009) hypothesised that adjustment size acts as a proxy for the significance and reliability of the received information. As a consequence, the FVA would be more likely to be improved from larger adjustments (though there is an obvious issue, in that small adjustments can inevitably only lead to small improvements in FVA). In addition, larger adjustments are likely to be made with greater consideration, as they can be more damaging if they are made in error. But the earlier published evidence has been limited. Here, we measure adjustment size as the percentage increase or decrease in the final forecast relative to the system forecast. This is a plausible measure in that a forecaster has the system forecast as an immediately salient reference point. An alternative measure is to examine the adjustments relative to the standard deviation of the past observations or the past system forecasts, which is smoother. Tables 8 and 9 summarise the effects of adjustment size in terms of the percentage of forecasts improved and the size of the improvements, respectively. In both cases for reasons of space, we present the results summarised over individual business units/companies.

The results show that whatever the size of the adjustments, the majority of positive adjustments led to negative FVA and worse bias, while negative adjustments (apart from the extreme of a final forecast of zero) had beneficial consequences. The beneficial effects of negative adjustments (Table 9) improved with the size of adjustment, and 67.4% of the larger adjustments resulted in improved accuracy. Bias is lessened with adjustment size for negative adjustments (again apart from the small number of zero final forecasts), though positive adjustments tend to increase bias, particularly for the largest adjustments which are unbounded. While these characterize the 'typical' company/business unit, the variability is high, with for example the retailer (dataset 2) performing poorly whatever the adjustment direction, and the

[10] While the datasets contain 22 business units/ companies in total, for dataset 4 we only include the three companies in the calculations, while in dataset 3 we exclude business unit 7 as previously noted due to its small number of observations.

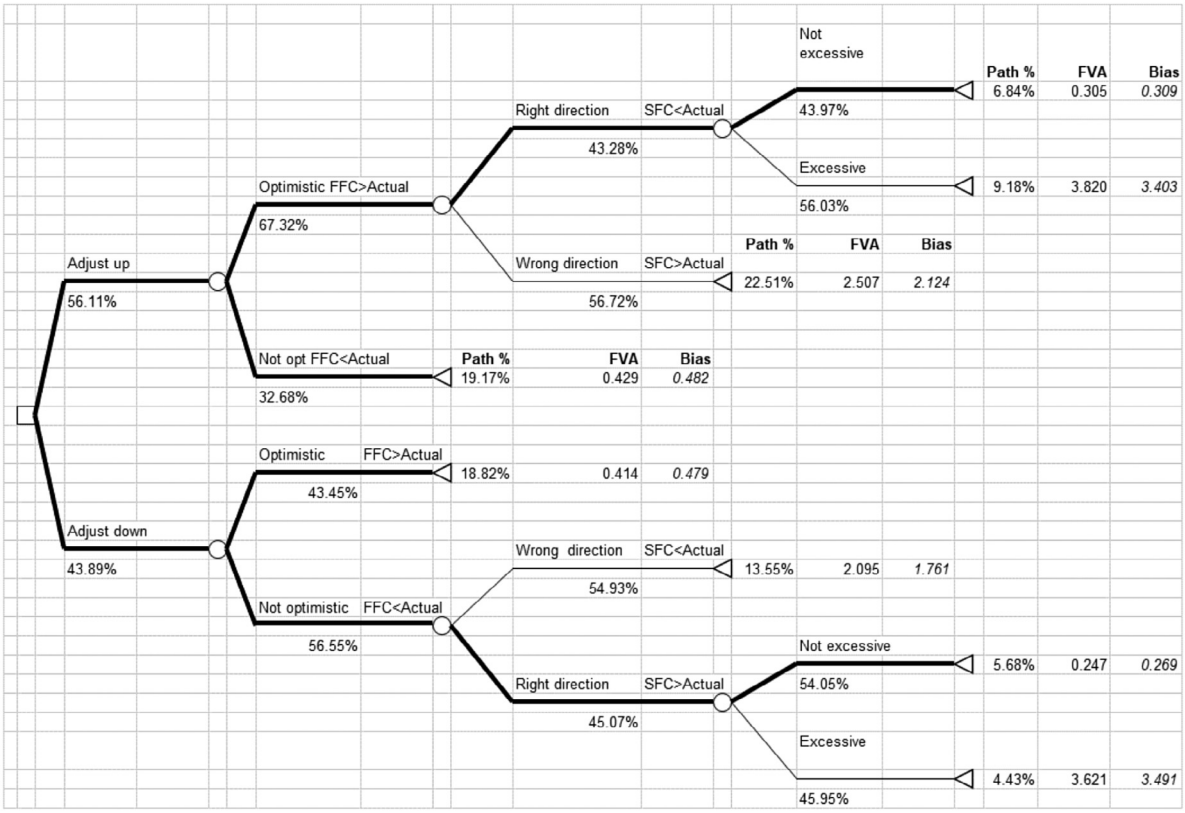

Fig. 1. Circumstances leading to beneficial or damaging FVA. (All figures are medians across the 16 business units.[10] FVA and bias, values below 1 indicate beneficial adjustments.)
N.B. FVA is measured using FVA2. Pathways leading to a positive FVA are shown in bold. Bias figures are shown in italics.

Table 7
Consistency of FVA and decision–outcome combinations across the 16 business units.

Decision	Outcome	% of adjusts on path		FVA2	
		Median	IQR	Median	IQR
Adjust up	Not excessive	6.84	3.71	0.31	0.07
	Excessive	9.18	3.93	3.82	0.51
	Wrong direction	22.51	5.33	2.51	0.42
	Not optimistic	19.17	4.79	0.43	0.07
Adjust down	Optimistic	18.82	4.06	0.41	0.06
	Wrong direction	13.55	2.35	2.10	0.21
	Not excessive	4.43	1.26	0.25	0.06
	Excessive	5.68	2.38	3.62	0.64

Table 8
Percentage of observations with positive FVA2 and bias reduction by adjustment size.

Adjustment direction		Adjustment size %					
		<10	10 to <50	50 to <100	100 to <250	≥250	Overall % adjustments with improved FVA2
Positive	% where FVA improved	46.5	43.1	47.4	49.1	56.0	45.9
	% where bias reduced	43.4	46.6	33.6	33.8	24.2	
	Median % of observations	31.6	43.8	11.3	7.1	3.1	
Negative	% where FVA improved	58.1	65.5	71.5	40.6		62.7
	% where bias reduced	51.8	56.8	43.5	40.6		
	Median % of observations	38.1	51.0	8.6	5.5		

Figures are medians across business units/companies.
* Negative adjustments are bounded by 100% as the maximum possible, equivalent to setting the final forecast to zero.

Table 9
Overall performance (bias and FVA) measured by medians across the business units.

Adjustment direction		Adjustment size %					Total no. of observations
		<10	10–50	50–100	100–250	>250	
Positive	FVA1	1.008	1.085	1.086	1.217	1.022	
	FVA2	1.040	1.132	1.235	1.396	1.314	
	Bias	1.052	1.129	1.051	1.173	0.797	
	Median % of observations	31.6	43.8	12.2	7.5	3.0	45 637
Negative	FVA1	0.951	0.798	0.489	1.389		
	FVA2	0.961	0.768	0.459	1.049		
	Bias	0.907	0.624	0.414	1.683		
	Median % of observations	37.9	46.7	8.4	4.7		38 833

Note that for negative information, a 100% adjustment is equivalent to setting the final forecast to zero. The percentages falling into the different adjustment classes are similar whether medians across business units/companies are used or the total percentage of observations.

business unit in dataset 4.2 performing well for positive adjustments yet badly for negative ones.

The results partially confirm earlier speculation: for positive information, large 'excessive' adjustments (of which there are relatively few) have negative FVA, while for negative adjustments, the larger the adjustment, the greater the benefit both with bias and FVA. The exception is those few negative adjustments where the final forecast is set equal to zero (see the 100–250 column). This pattern proved consistent across datasets, with only one unit showing improvement with size for positive adjustments. Appendix A.3 gives details of variations in the effect of the size and direction of adjustments on FVA for the individual datasets.

8. Is it possible to improve accuracy by debiasing judgmental adjustments?

In the previous sections, we saw how various theoretically non-diagnostic cues consistently affect the adjustment direction and size. We now ask whether modifying the adjustment to remove the effect of these cues will increase forecast accuracy. The models we propose are two forms of an error bootstrap model (Fildes & Hastings, 1994): an 'optimal weight' model, and a full model. We first estimate these models on approximately 75% of the data and then test them on the remainder.

The process by which a forecaster adjusts is that the 'advice' of the system forecast, SFC_t is considered and then an adjustment decided on, Adj_t. Effectively a weight is given to the system forecast and the judgment. This raises the question of whether the weight is appropriate or whether a better weighting scheme is conceptually available to the forecaster. The error then depends on these two factors and their weights.

A model of the error can suggest a reweighting, such as Blattberg and Hoch's (1990) 50–50 scheme where the model and a judgmental forecast are given equal weights. Translating the Blattberg and Hoch scheme to the system forecast, the adjustment effectively damps the adjustment by 50%. This was shown to be effective by Franses and Legerstee (2009) and Baecke et al. (2017), though the latter noted that it also reduced the effects of adjustments that were beneficial. Fildes et al. (2009) demonstrated that this 50–50 weighting is not always optimal, particularly for positive adjustments. An optimal weighting scheme can be estimated through the 'optimal weight' model,[11] which asks the question whether, post hoc, a weighting could be found which would outperform the forecasters' adjustment:

$$Final\ Forecast\ Error_t = (intercept) + \beta_1 SFC_t + \beta_2 ADj_t$$

[Optimal weight model]

The variables, as above, were normalised by the standard deviation of the system forecast. The inclusion of an intercept provides a descriptive model, with the intercept representing (normalised) consistent bias while its exclusion suggests an easily interpretable reweighting scheme. Without the intercept, this permits the forecaster in principle to reweight the adjustment and the system forecast. The interpretation is as follows: with an estimated parameter of β for either variable X, adding in βX leads to an improved forecast. With β_1 or β_2 non-zero, reweighting the system forecast and/or the judgmental adjustment would in principle lead to a smaller error and an improved final forecast. The β_2 term could potentially be dependent on the system forecast, which would lead to an interaction term, distinguishing the model form from a straightforward combination.

A second approach is to use the full model discussed in the previous section, which includes an extended number of observed cues as follows:

$Final\ Forecast\ Error_t = f\ (ADJ_t, System\ Forecast_t, Previous\ adjustment, Change_t, Previous\ final\ forecast\ error_{t-1};)$ [Full Model]

The models are estimated for each company/business unit on approximately 75% of the early data in the whole dataset and the accuracy measures calculated on the remainder, a stringent test in that the test data are outside the training period.[12] The medians estimated for

[11] Our objective is to forecast the underlying signal of the time series, but given that it is unobserved, we can never be sure whether our forecast was truly 'optimal' apart from noise. However, our accuracy measures are averaged over a large number of periods and cases, so we can assess which approaches are likely to be closer to optimal.

[12] Dataset 4 is cross-sectional, so no lagged variables are available. A 75%–25% training–test sample was used.

Table 10
Test data accuracy of model adjusted forecasts: medians over all companies/business units.[a]

Adjust	Accuracy measures	System forecast (SFC)	Final forecast	50–50	Optimal weight of SFC and Adj	Full model
Up	MdAPE	24.9%	28.4%	22.3%	21.8%	20.0%
	MdFVA2		1.060	0.941	0.969	0.811
	GMFVA2		1.089	0.933	0.927	0.811
Down	MdAPE	23.2%	20.0%	21.7%	20.4%	16.2%
	MdFVA2[b]		0.869	0.847	0.803	0.783
	GMFVA2		0.830	0.868	0.788	0.731

[a] As usual, business unit 7, dataset 3 was excluded in the calculations.
[b] Using a geometric mean gives a similar result.

the optimal weight of the system forecast (*SFC*) and the adjustment (*Adj*) are as follows:

Adjustment direction	SFC	Adj
Up	−0.033	−0.598
Down	−0.019	−0.598

The signs are consistent across business units, with the headline conclusion that too much weight is given to the judgmental adjustment (some 90% have a negative sign) and this is particularly true for upward adjustments. In addition, the upward forecasts are overweighted. The patterns are very consistent: for both the positive and negative adjustments, the weight given to the system forecast is close to optimal (though sometimes significant) while the adjustment is overweighted particularly for positive adjustments, with median overweighting of 0.6.

The results showing the relative accuracy of the different error models are summarised in Table 10. (The estimates for the individual business units/companies are given in the online spreadsheet *Regmodels*.) As Davydenko and Fildes (2013) argued, the statistical characteristics of the GMFVA are preferable to other measures and may tell different stories. Here we present three measures in Table 10 to summarise the results: the MAPE, MdFVA2, and geometric mean FVA2.

The full model generally performs better than the optimal weighting model, although the picture is more nuanced, depending in part on the error measure and the individual business units and companies. For example, dataset 6 shows major improvements while for dataset 3 there is no gain. The simple reweighting scheme is itself approximately 17% better than the final adjusted forecasts for upward adjustments, and 6% for downward ones. The 50–50 scheme is often a poor competitor compared to the full model. The differences between companies/business units emphasises the different cues in both the demand planning system and the forecasting software that affect the demand planner's judgment.

This is an out-of-sample comparison based on one-period-ahead errors. A complementary way of evaluating these compensation models is to use different data windows for estimating the models and a corresponding test set for validating the relative accuracy statistics. We had data limitations arising from the limited lengths of the individual series but chose to use a minimum of four different forecast origins and to evaluate over the remainder of the time periods (apart from dataset 5, which is too limited). The results mirror the figures given in Table 10 and are available in the online spreadsheet *RegModels*.

The picture we get from this analysis of whether there are consistent errors in the adjustment process is that by removing those observable determinants of the adjustments, identified in section 6, we are able to consistently improve accuracy and hence FVA.

9. Conclusions

9.1. Discussion and implications for practice

In demand planning, the operational system forecasts are derived from a diverse set of sources which are integrated into a final forecast and used throughout the organization but particularly in the supply chain. Our objective was to understand how demand planners integrate this diverse information and its effects on forecast accuracy and bias—forecast value added. Although our data were gathered from companies operating in different countries and conditions, many of our findings revealed common patterns in decisions to adjust system forecasts and their effect.

Overall, our results indicated that only half of adjustments succeeded at improving forecast accuracy and bias. However, consistent with earlier findings, downward adjustments consistently improved accuracy and bias, sometimes substantially. In contrast, upward adjustments had a mixed record and often appreciably damaged accuracy, primarily because they were excessive, suggesting that optimism bias or motivations to inflate forecasts were widespread. The decision to adjust system-based forecasts is usually justified when the forecaster has access to important information not available to the system's algorithm, such as special events or unusual market conditions. We found evidence that forecasters had some ability to detect when these factors suggest that an intervention is required. For example, they recognised the appropriate direction of adjustment in 59.9% of cases when adjusting upwards, and 69.1% when adjusting downwards. However, it appears that they were unable to exploit such information fully. Mostly, they were distracted by non- or less diagnostic cues that are typically displayed in forecasting support systems, such as the previous forecast error or the previous adjustment. Some of these factors, particularly whether there was a previous

adjustment, had a significant effect on decisions to adjust. This is consistent with a laboratory study by Fildes et al. (2019), who found that salient, but non-diagnostic, information was strongly associated with adjustment decisions. These findings have potential implications for the design of forecasting support systems and the prominence with which they display information. For example, software suppliers have tended to increase the volume and complexity of information available to forecasters, and our analysis suggests this may be counterproductive. They also raise a key question: Why, with such emphasis on the sales and operations planning process, is there a failure to effectively incorporate key features of the developing business environment into the demand plans?

A robust finding was that adjustments tended to be inefficient. Damping them by reducing their weighting when combining them with the system forecast would have led to improved accuracy, confirming previous findings (e.g. Baecke et al., 2017; Franses & Legerstee, 2009). Even greater accuracy could have been obtained by forecasting the error arising from adjustments based on cue variables such as the previous observation, the system forecast, and the previous adjustment, and using this to correct the final forecast. The potential improvements from using the full model proposed above are substantial: around 25% for upward adjustments, and 10% for downward adjustments, 5% compared to the adjusted forecast (Table 10). This presents a research challenge with important implications for practice: to design organizational processes and a FSS that capitalises on these sub-optimal weightings and refocuses the forecasters on the new information relevant to the future demand. This is of increasing importance, as machine learning methods, which attempt to capture more features of the determinants of demand, have become increasingly integrated into demand planning systems. Our findings indicate a dire need for the academic field and practitioners to work closely together to solve the issue of inefficient judgmental adjustment and heighten overall forecast accuracy.

Methodologically, our study has shown the importance of analysing a variety of datasets using modelling methods (e.g. cross-validation) to establish robust findings across organizations and business units but also the extent of diversity and the characteristics of cases where adjustment tends to be more successful. A synthesis of field studies, experiments, and case studies is required to provide the depth of understanding needed to underpin improvements in forecast value added. The current primary reliance on experiments (Perera et al., 2019) can only point to possible ways forward.

9.2. Limitations and future research

Our research involved a wide range of companies and applied an innovative set of analyses to over 400,000 forecasts, which when cleaned left around 145,000 for modelling. As such, it is the most extensive study to date to explore the important issue of how people interact with system forecasts. However, inevitably our findings have a number of limitations. We were not privy to the organizational processes that produced the final forecasts (Fildes et al., 2019) or the complete sets of information that forecasters had access to. We therefore acknowledge that our models do not include the many organizational and market factors that may also have influenced adjustment behaviour (see for example the process discussions in Fildes & Goodwin, 2021). For example, Fildes et al. (2009) and Fahimnia et al. (2022) demonstrated that service levels can have an influence on adjustment behaviour. Nor were we aware of the forecasters' motivations, which may have included objectives other than improved accuracy. However, surveys have suggested that accuracy is the prime objective in demand forecasting (e.g. Fildes & Goodwin, 2007) and, irrespective of an individual forecaster's motivations, users of forecasts are likely to make their decisions assuming that the forecasts represent true and honest expectations of future demand. Nevertheless, more organizational field studies are needed to gain insight into processes and motivations. Additionally, our analysis was limited to one-period-ahead forecasts. Other studies (e.g. Van den Broeke et al., 2019) have found that the forecast lead time can have a significant effect on adjustment behaviour and accuracy. Future researchers could therefore usefully apply our methods to forecasts over multiple lead times. In general, there is a significant need for more research with company data to test the generalizability of the reported findings, and to dig deeper into the cause of variations in adjustment efficiency between companies, as highlighted above. Some of the more curious results, which raise theoretical questions, need to be confirmed or contradicted in experimental studies and further investigated in field research. Methodologically there are a number of approaches to analysing multiple datasets which might usefully provide different perspectives.

Finally, our databases spanned years during which methods embodied in forecasting software continued to develop (Goodwin, Hoover, Makridakis, Petropoulos, & Tashman, 2023). In particular, more complex black-box machine learning methods have become increasingly common (Schaer, Sventunkov, Yusupova, & Fildes, 2022) and it is possible that forecaster adjustment behaviour may also be changing in response to these developments.

Given the near ubiquity of human interventions where computer-based forecasts are available, an understanding of the factors underlying adjustments and their value is needed. Despite the limitations of our study, we hope that it will contribute to this and stimulate further research that is firmly based in the fast-developing organizational processes stimulated by data science. This in turn should support enhancements to the design of forecasting support systems and lead to a more efficient and effective use of forecasters' time so that the forecasts they produce are less biased and more accurate. The practical results would be improved performance and less waste.

CRediT authorship contribution statement

Robert Fildes: Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Paul Goodwin:** Writing – review & editing, Writing – original draft, Validation, Methodology, Formal analysis, Conceptualization. **Shari De Baets:** Writing – review & editing, Writing – original draft, Methodology.

Table A.1
Commentary on individual datasets.

Set 1	Consistent performance across companies. While the previous adjustment is always significant and substantial, only the positive and negative information is consistent.
Set 2	Little is gained by modelling each supplier separately. SKUs are not significant, the key variables being the previous adjustment and the system forecast.
Set 3	Broadly consistent performance across business units though the paucity of non-adjusted data makes some of the estimates unreliable—business unit 7 in particular. The estimates used in Table 5 are the medians across just the other six business units.
Set 5	Only a previous negative adjustment has a substantial impact on there being an adjustment in the following period. The change and system forecast error are significant.
Set 6	Here all variables are significant in certain circumstances, with the corresponding direction of a previous adjustment most impactful. In addition, extreme information in all the variables is significant. However, there is a limited effect on in-sample overall accuracy. With only a few unadjusted observations, the model is unable to identify them.

Table A.2
Distribution statistics for FVA2: Selected datasets.

Selected	Company	Info	nobs	Geometric Mean FVA2	MdFVA2	qrange	p10FVA2	p90FVA2	p5FVA2	p95FVA2
data set 1	1	1	1157	0.75	0.83	1.38	0.12	3.99	0.06	7.34
		−1	1086	0.70	0.71	1.05	0.13	3.87	0.06	7.80
	3	1	1155	1.02	0.93	1.84	0.19	6.00	0.10	11.95
		−1	510	0.69	0.75	0.98	0.13	3.05	0.08	6.15
data set 4	3	1	318	0.33	0.33	0.54	0.07	1.54	0.04	2.34
		−1	250	0.33	0.40	1.10	0.02	3.51	0.01	7.51
data set 6		1	1807	0.50	0.67	1.20	0.06	3.00	0.03	4.86
		−1	1953	0.74	0.70	0.87	0.25	2.50	0.17	4.00

Notes. Info: 1 and −1 imply positive and negative information, respectively. 'nobs' = number of observations. qrange is the interquartile range while p10 etc. are the percentiles.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Robert Fildes reports financial support, equipment, drugs, or supplies, and travel were provided by Lancaster University Department of Management Science. Robert Fildes reports a relationship with Lancaster University Management School that includes: non-financial support and travel reimbursement. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Our thanks are due to the providers of the datasets used and the additional information needed for their interpretation.

Appendix A

A.1. Commentary on individual datasets

See Table A.1.

A.2. Distribution of FVA

The distributions of FVA2 (used because it is associated with each observation), presented log transformed, are displayed in Fig. A.1 for negative and positive information. The distribution of FVA is of potential importance in that it illustrates the risk incurred when adjustments are made, and it has been overlooked in previous studies. It is illustrated for four different datasets chosen to show differences between companies/business units.

Overall, Fig. A.1 demonstrates the near log-normality of forecast value added, apart from the tails. Note that a negative value of logFVA2 shows an improvement due to adjustment. The four graphs show adjustments differing in their medians for different directions of adjustment, from no difference between their location and shape (dataset 1: company 1) to positive adjustments tending to both add value (dataset 6) and worsen value (dataset 1: company 3, dataset 4). Also, both skew and kurtosis differ, sometimes substantially (dataset 6).

Table A.2 shows distribution statistics for the selected datasets. The percentiles and associated ranges underline substantial differences between dataset 4, company 3 that had effective interventions (both positive and negative) and one like dataset 1, company 3 with a wide range and extreme tail for negative FVAs. However, it is important to note that the distribution examples shown above were chosen to illustrate the differences in FVA rather than the commonalities.

A.3. Effects of size and direction at company/business unit level on FVA

See Table A.3.

Table A.3
Effects of size and direction at company/business unit level on FVA.

Dataset 1	Overall FVA <1, negative information improves with adjustment size for all companies, apart from business unit 2, where setting the FFC to zero has seriously negative consequences; positive adjustments typically make things worse with increased size.
Dataset 2	Extremely damaging effects when FFC set to zero. Fildes et al. (2009) speculated that inventory control considerations had influenced these adjustments.
Dataset 3	FVA for positive information poor. Improves with size for negative adjustments, as does bias. No final forecasts were set to zero.
Dataset 4	For positive adjustments, companies not uniform, nor product groups within companies. E.g. company 2, business unit 1 adjustment improves both bias and FVA, apart from the smallest. Negative adjustments improve with size apart from those setting final to 0. Company 3 excellent, apart from when FFC set to 0.
Dataset 5	Conforms to the overall pattern, with negative adjustments improving FVA and bias as size increases
Dataset 6	Overall positive adjustments add value and lower bias, but these are due to the substantial number of small adjustments. Negative adjustments, apart from a final forecast of 0, improve FVA and bias, improving with size.

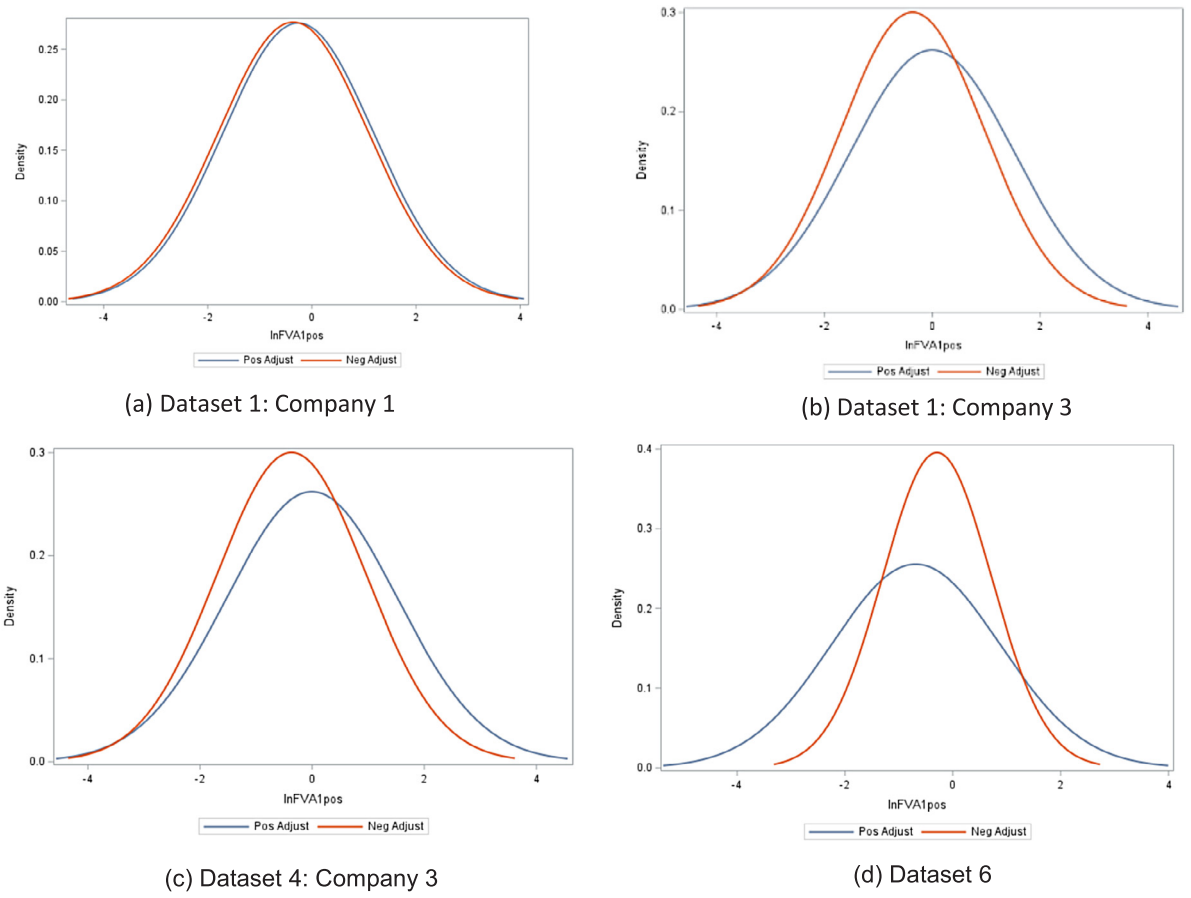

Fig. A.1. Distribution of log forecast value added (logFVA2) for four illustrative business units/companies contrasting positive and negative information.

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ijforecast.2024.07.006.

References

Alvarado-Valencia, J., Barrero, L. H., Önkal, D., & Dennerlein, J. T. (2017). Expertise, credibility of system forecasts and integration methods in judgmental demand forecasting. *International Journal of Forecasting*, *33*, 298–313.

Andreassen, P. B., & Kraus, S. J. (1990). Judgmental extrapolation and the salience of change. *Journal of Forecasting*, *9*, 347–372.

Asimakopoulos, S., Dix, A., & Fildes, R. (2011). Using hierarchical task decomposition as a grammar to map actions in context: Application to forecasting systems in supply chain planning. *International Journal of Human-Computer Studies*, *69*, 234–250.

Baecke, P., De Baets, S., & Vanderheyden, K. (2017). Investigating the added value of integrating human judgement into statistical

demand forecasting systems. *International Journal of Production Economics*, *191*, 85–96.
Baker, J. (2021). Maximizing forecast value added through machine learning and nudges. *Foresight: The International Journal of Applied Forecasting*, *60*, 8–15.
Blattberg, R. C., & Hoch, S. J. (1990). Database models and managerial intuition: 50-% model + 50-% manager. *Management Science*, *36*, 887–899.
Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision making: An integrative literature review, and implications for the organizational science. *Organizational Behavior and Human Decision Processes*, *101*, 127–151.
Carbone, R., Andersen, A., Corriveau, Y., & Corson, P. P. (1983). Comparing for different time series methods the value of technical expertise individualized analysis, and judgmental adjustment. *Management Science*, *29*, 559–566.
Davydenko, A., & Fildes, R. (2013). Measuring forecasting accuracy: The case of judgmental adjustments to SKU-level demand forecasts. *International Journal of Forecasting*, *29*, 510–522.
Davydenko, A., & Fildes, R. (2015). Forecast error measures: Critical review and practical recommendations. In M. Gilliland, L. Tashman, & U. Sglavo (Eds.), *Business Forecasting* (pp. 238–258). Hoboken, New Jersey: Wiley.
Davydenko, A., & Goodwin, P. (2021). Assessing point forecast bias across multiple time series: Measures and visual tools. *International Journal of Statistics and Probability*, *10*, 46–69.
De Baets, S., & Harvey, N. (2020). Using judgment to select and adjust forecasts from statistical models. *European Journal of Operational Research*, *284*, 882–895.
De Baets, S., & Harvey, N. (2023). Incorporating external factors into time series forecasts. In M. Seifert (Ed.), *Judgment in Predictive Analytics*. New York: Springer.
Diamantopoulos, A., & Mathews, B. P. (1989). Factors affecting the nature and effectiveness of subjective revision in sales forecasting: An empirical study. *Managerial and Decision Economics*, *10*, 51–59.
Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*, 114–126.
Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, *64*, 1155–1170.
Einhorn, H. J. (1974). Expert judgment: Some necessary conditions and an example. *Journal of Applied Psychology*, *59*(562).
Eroglu, C., & Croxton, K. L. (2010). Biases in judgmental adjustments of statistical forecasts: The role of individual differences. *International Journal of Forecasting*, *26*, 116–133.
Eroglu, C., & Sanders, N. R. (2021). Effects of personality on the efficacy of judgmental adjustments of statistical forecasts. *Management Decision*, *60*, 589–605.
Fahimnia, B., Arvan, M., Tan, T., & Siemsen, E. (2022). A hidden anchor: The influence of service levels on demand forecasts. *Journal of Operations Management*, *69*, 856–871.
Fildes, R. (1991). Efficient use of information in the formation of subjective industry forecasts. *Journal of Forecasting*, *10*(6), 597–617.
Fildes, R., & Goodwin, P. (2007). Against your better judgment? How organizations can improve their use of management judgment in forecasting. *Interfaces*, *37*, 570–576.
Fildes, R., & Goodwin, P. (2021). Stability in the inefficient use of forecasting systems: A case study in a supply chain company. *International Journal of Forecasting*, *37*, 1031–1046.
Fildes, R., Goodwin, P., Lawrence, M., & Nikolopoulos, K. (2009). Effective forecasting and judgmental adjustments: An empirical evaluation and strategies for improvement in supply-chain planning. *International Journal of Forecasting*, *25*, 3–23.
Fildes, R., Goodwin, P., & Onkal, D. (2019). Use and misuse of information in supply chain forecasting of promotion effects. *International Journal of Forecasting*, *35*, 144–156.
Fildes, R., & Hastings, R. (1994). The organization and improvement of market forecasting. *Journal of the Operational Research Society*, *45*, 1–16.
Fildes, R., & Petropoulos, F. (2015). Improving forecast quality in practice. *Foresight: International Journal of Applied Forecasting*, *36*, 5–12.
Franses, P. H., & Legerstee, R. (2009). Properties of expert adjustments on model-based SKU-level forecasts. *International Journal of Forecasting*, *25*, 35–47.
Franses, P. H., & Legerstee, R. (2010). Do experts' adjustments on model-based SKU-level forecasts improve forecast quality? *Journal of Forecasting*, *29*(3), 331–340.
Franses, P. H., & Legerstee, R. (2011). Experts' adjustment to model-based SKU-level forecasts: Does the forecast horizon matter? *Journal of the Operational Research Society*, *62*(3), 537–543.
Gal, D. (2006). A psychological law of inertia and the illusion of loss aversion. *Judgment and Decision Making*, *1*, 23–32.
Galbraith, C. S., & Merrill, G. B. (1996). The politics of forecasting: Managing the truth. *California Management Review*, *38*, 29–43.
Gigerenzer, G., Todd, P. M., & the ABC Research Group (1999). *Simple Heuristics that Make Us Smart*. New York: Oxford University Press.
Gilliland, M. (2008). Forecast value added analysis: Step-by-step. SAS White Paper.
Goodwin, P., & Fildes, R. (1999). Judgmental forecasts of time series affected by special events: Does providing a statistical forecast improve accuracy? *Journal of Behavioral Decision Making*, *12*, 37–53.
Goodwin, P., Fildes, R., Lawrence, M., & Nikolopoulos, K. (2007). The process of using a forecasting support system. *International Journal of Forecasting*, *23*, 391–404.
Goodwin, P., Hoover, J., Makridakis, S., Petropoulos, F., & Tashman, L. (2023). Business forecasting methods: Impressive advances, lagging implementation. *Plos One*, *18*(12), Article e0295693.
Harvey, N., Ewart, T., & West, R. (1997). Effects of data noise on statistical judgement. *Thinking & Reasoning*, *3*(2), 111–132.
Jarvenpaa, S. L. (1990). Graphic displays in decision making—The visual salience effect. *Journal of Behavioral Decision Making*, *3*(4), 247–262.
Karelse, J. (2021). Mitigating unconscious bias in forecasting. *Foresight: The International Journal of Applied Forecasting*, *61*, 5–14.
Khosrowabadi, N., Hoberg, K., & Imdahl, C. (2022). Evaluating human behaviour in response to AI recommendations for judgemental forecasting. *European Journal of Operational Research*, *303*, 1151–1167.
Kourentzes, N., & Fildes, R. (2023). The dynamics of judgemental adjustments in demand planning. SSRN Working Paper.
Kremer, M., Moritz, B., & Siemsen, E. (2011). Demand forecasting behavior: System neglect and change detection. *Management Science*, *57*, 1827–1843.
Krizan, Z., & Windschitl, P. D. (2007). The influence of outcome desirability on optimism. *Psychological Bulletin*, *133*(95).
Lawrence, M., Goodwin, P., O'Connor, M., & Onkal, D. (2006). Judgmental forecasting: A review of progress over the last 25 years. *International Journal of Forecasting*, *22*, 493–518.
Lim, J. S., & O'Connor, M. (1996). Judgmental forecasting with time series and causal information. *International Journal of Forecasting*, *12*, 139–153.
Massey, C., & Wu, G. (2005). Detecting regime shifts: The causes of under- and overreaction. *Management Science*, *51*, 932–947.
Mathews, B. P., & Diamantopoulos, A. (1986). Managerial intervention in forecasting: An empirical investigation of forecast manipulation. *International Journal of Research in Marketing*, *3*, 3–10.
Mathews, B. P., & Diamantopoulos, A. (1990). Judgmental revision of sales forecasts : Effectiveness of forecast selection. *Journal of Forecasting*, *9*, 407–415.
Mathews, B. P., & Diamantopoulos, A. (1992). Judgmental revision of sales forecasts - the relative performance of judgementally revised versus nonrevised forecasts. *Journal of Forecasting*, *11*, 569–576.
Mello, J. (2009). The impact of sales forecast game playing on supply chains. *Foresight: The International Journal of Applied Forecasting*, *1*, 3–22.
Moritz, B., Siemsen, E., & Kremer, M. (2014). Judgmental forecasting: Cognitive reflection and decision speed. *Production and Operations Management*, *23*(7), 1146–1160.
Oliva, R., & Watson, N. (2009). Managing functional biases in organizational forecasts: A case study of consensus forecasting in supply chain planning. *Production and Operations Management*, *18*, 138–151.
Perera, H. N., Hurley, J., Fahimnia, B., & Reisi, M. (2019). The human factor in supply chain forecasting: A systematic review. *European Journal of Operational Research*, *274*, 574–600.

Petropoulos, F., Fildes, R., & Goodwin, P. (2016). Do 'big losses' in judgmental adjustments to statistical forecasts affect experts' behaviour? *European Journal of Operational Research*, *249*, 842–852.

Platzer, C., & Bröder, A. (2012). Most people do not ignore salient invalid cues in memory-based decisions. *Psychonomic Bulletin & Review*, *19*(4), 654–661.

Platzer, C., Bröder, A., & Heck, D. W. (2014). Deciding with the eye: How the visually manipulated accessibility of information in memory influences decision behavior. *Memory & Cognition*, *42*, 595–608.

Remus, W., O'Connor, M., & Griggs, K. (1995). Does reliable information improve the accuracy of judgmental forecasts. *International Journal of Forecasting*, *11*, 285–293.

Sanders, N. R. (1992). Accuracy of judgmental forecasts: A comparison. *Omega*, *20*, 353–364.

Schaer, O., Sventunkov, I., Yusupova, A., & Fildes, R. (2022). Survey: Forecasting software trends in a challenging world. *OR/MS Today*, *49*(5).

Seifert, M., Siemsen, E., Hadida, A. L., & Eisingerich, A. B. (2015). Effective judgmental forecasting in the context of fashion products. *Journal of Operations Management*, *36*, 33–45.

Sroginis, A., Fildes, R., & Kourentzes, N. (2022). Use of contextual and model-based information in adjusting promotional forecasts. *European Journal of Operational Research*, *307*, 1177–1191.

Syntetos, A. A., Nikolopoulos, K., Boylan, J. E., Fildes, R., & Goodwin, P. (2009). The effects of integrating management judgement into intermittent demand forecasts. *International Journal of Production Economics*, *118*, 72–81.

Taleb, N. (2005). *The black swan: Why Don't We Learn That We Don't Learn*. NY: Random House.

Trapero, J. R., Pedregal, D. J., Fildes, R., & Kourentzes, N. (2013). Analysis of judgmental adjustments in the presence of promotions. *International Journal of Forecasting*, *29*, 234–243.

Van den Broeke, M., De Baets, S., Vereecke, A., Baecke, P., & Vanderheyden, K. (2019). Judgmental forecast adjustments over different time horizons. *Omega*, *87*, 34–45.

Webby, R., & O'Connor, M. (1996). Judgemental and statistical time series forecasting: A review of the literature. *International Journal of Forecasting*, *12*, 91–118.